

Inventa Analytics: criando vantagem competitiva de negócio com dados + IA

Lucas Fonseca

Agenda e Pontos-Chave

- Apresentação pessoal
- Escopo da conversa
- Fundação de Dados
- Revisão Conceitual
- Inventa Analytics
- Código
- Limitações e Próx. Passos



Prazer, Lucas

De Campos Novos, Cunha - SP

Formado em Análise e Desenvolvimento de Sistemas pela FATEC-SP em 2016

Especialista em Inteligência e gestão de dados pela Escola Politécnica da USP (2018)

Cursando Gestão Empresarial na EAESP - FGV (2025)

Há 12 anos no mercado de tecnologia. Experiência em Engenharia de SW e Dados, Analytics Engineering, Machine Learning Ops e liderança

Membro do AGI Club (by Ifood)



Lavandário • Cunha-SP



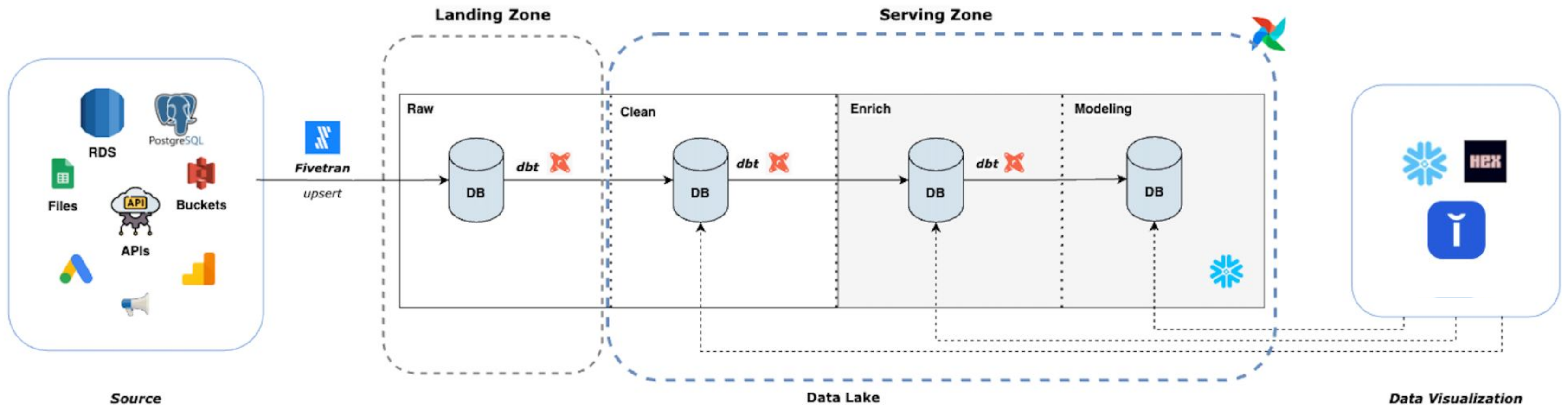
Escopo da conversa

- ❖ Conversaremos sobre como estamos construindo o **Inventa Analytics** , uma solução conversacional baseada em Dados+LLM para self-service em dados na Inventa (B2x Holding).
- ❖ Engenharia de dados como fundação para produtos de Dados e IA.
- ❖ Arquitetura modular de agentes para interpretação de intenções, geração queries e entrega de insights confiáveis.
- ❖ Tech Stack do projeto: Snowflake, OpenAI, FastAPI.
- ❖ Aprendizados em produtos inovadores, equilibrando confiabilidade e experimentação.
- ❖ Experiência do usuário (UX + AI) como parte central da arquitetura.
- ❖ Do PoC ao MVP até a evolução para produto em produção.

*Engenharia de dados como
fundação para produtos de
Dados e IA*

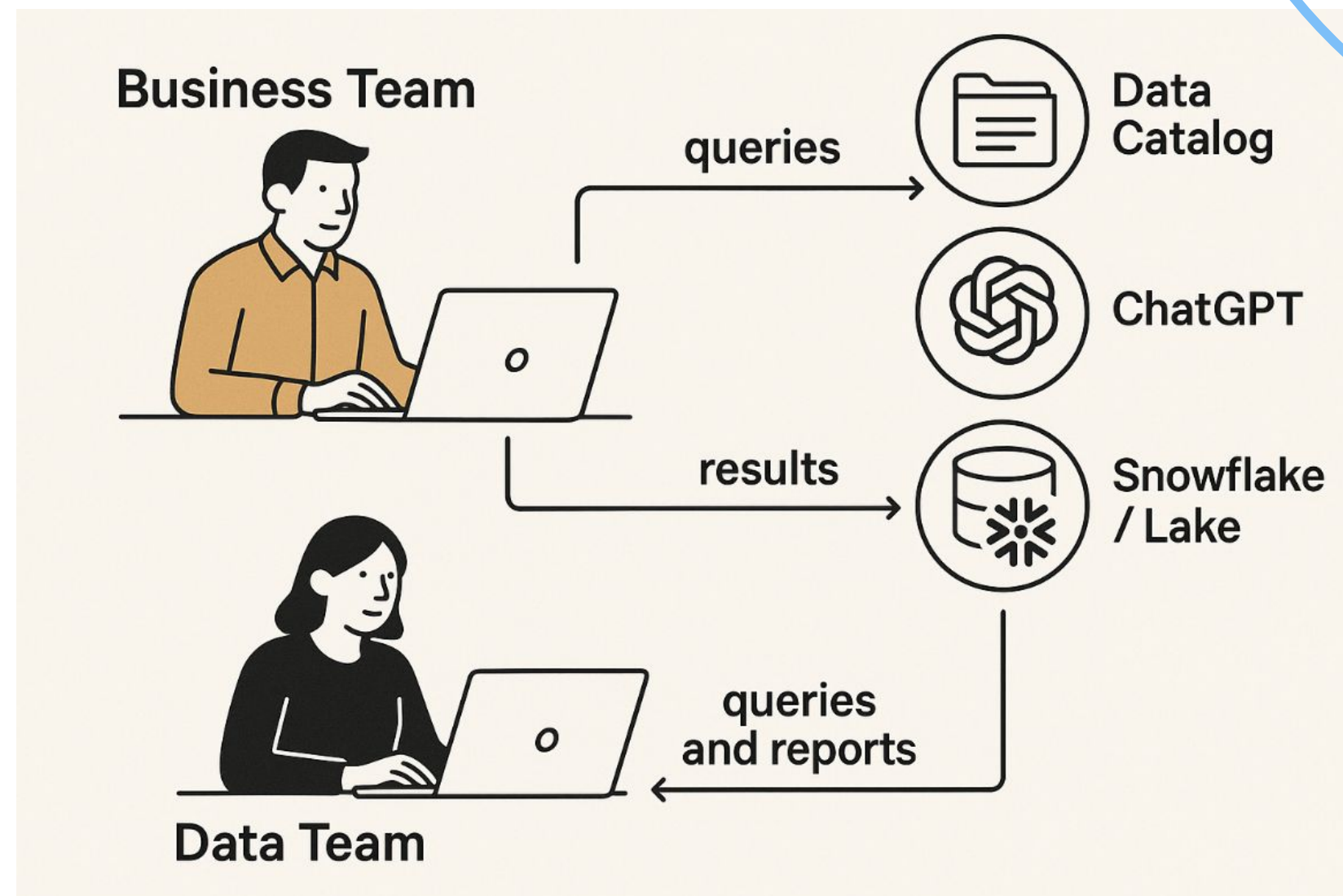


Artemis Data Platform



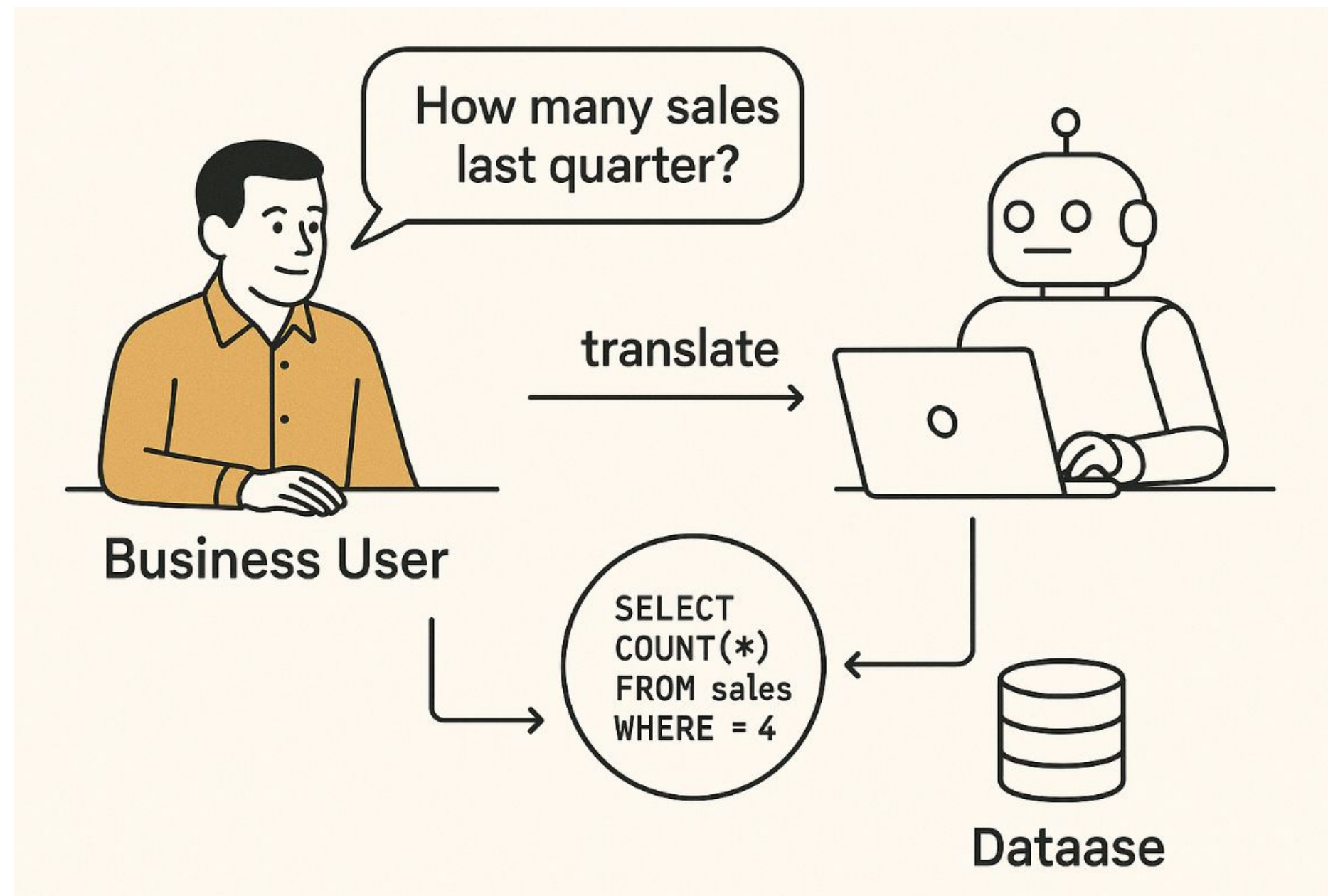
Problema

- ❖ Time de negócios não possui fluência em SQL para consultar os dados
- ❖ Time de Dados se tornou gargalo para a empresa
 - Manutenção da Plataforma
 - Criação de Produtos de Dados
 - Consultas de Dados

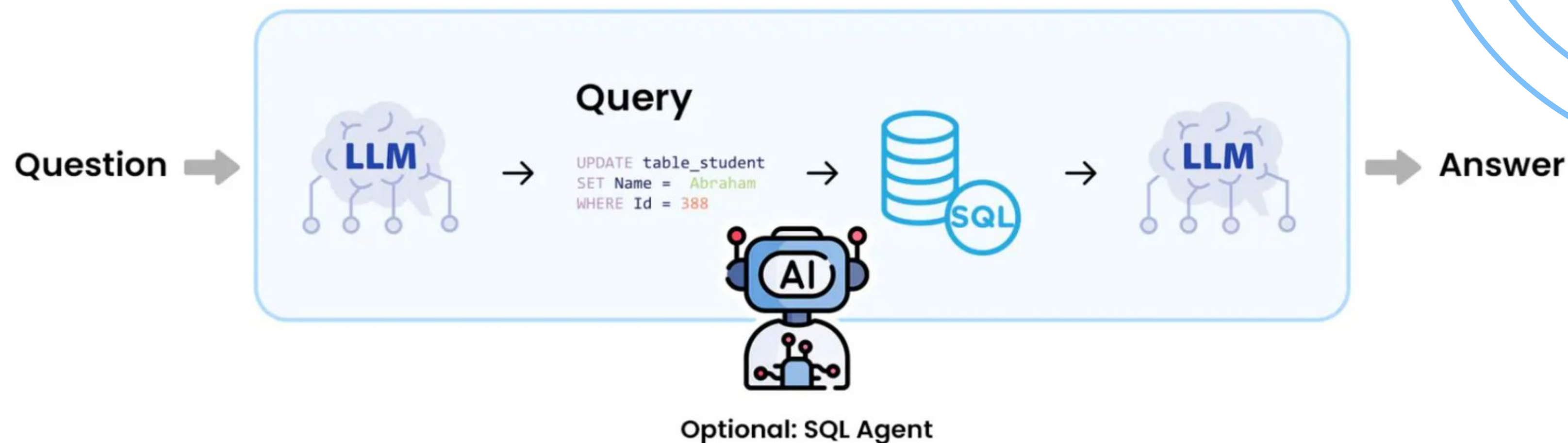


Problema

- ❖ Ponto central: Como poderíamos democratizar os dados da Inventa através de tecnologia?



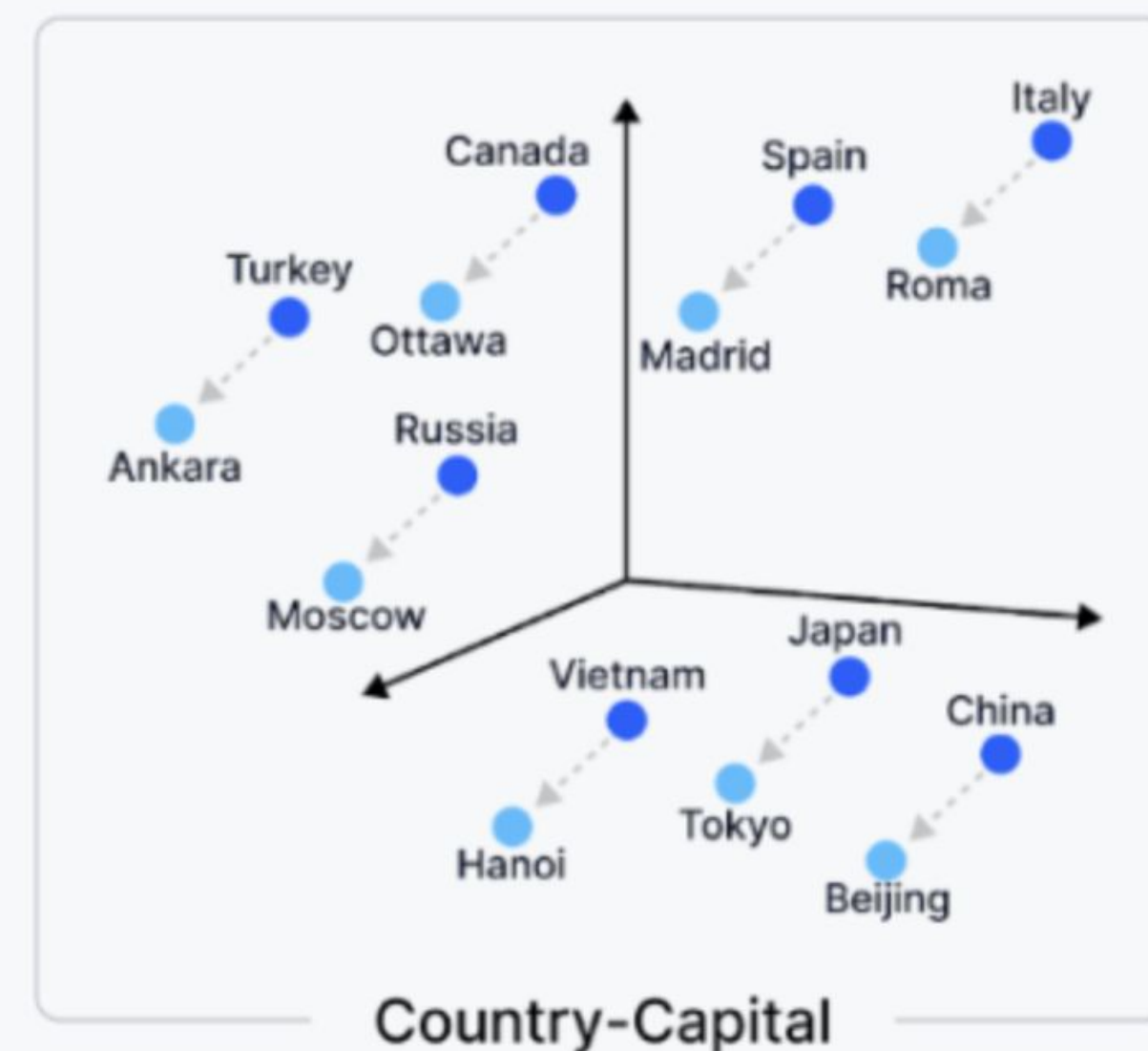
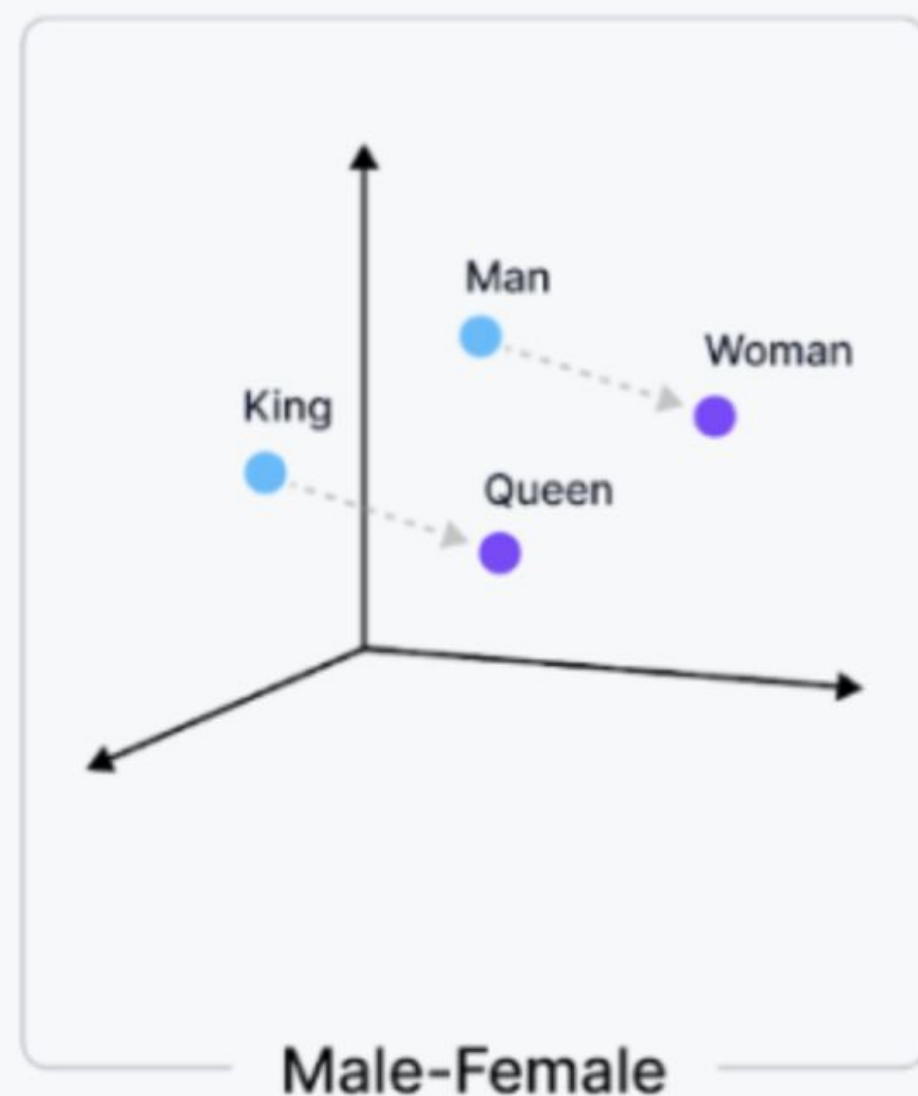
Text-to-Sql / NL2SQL



Algumas definições conceituais

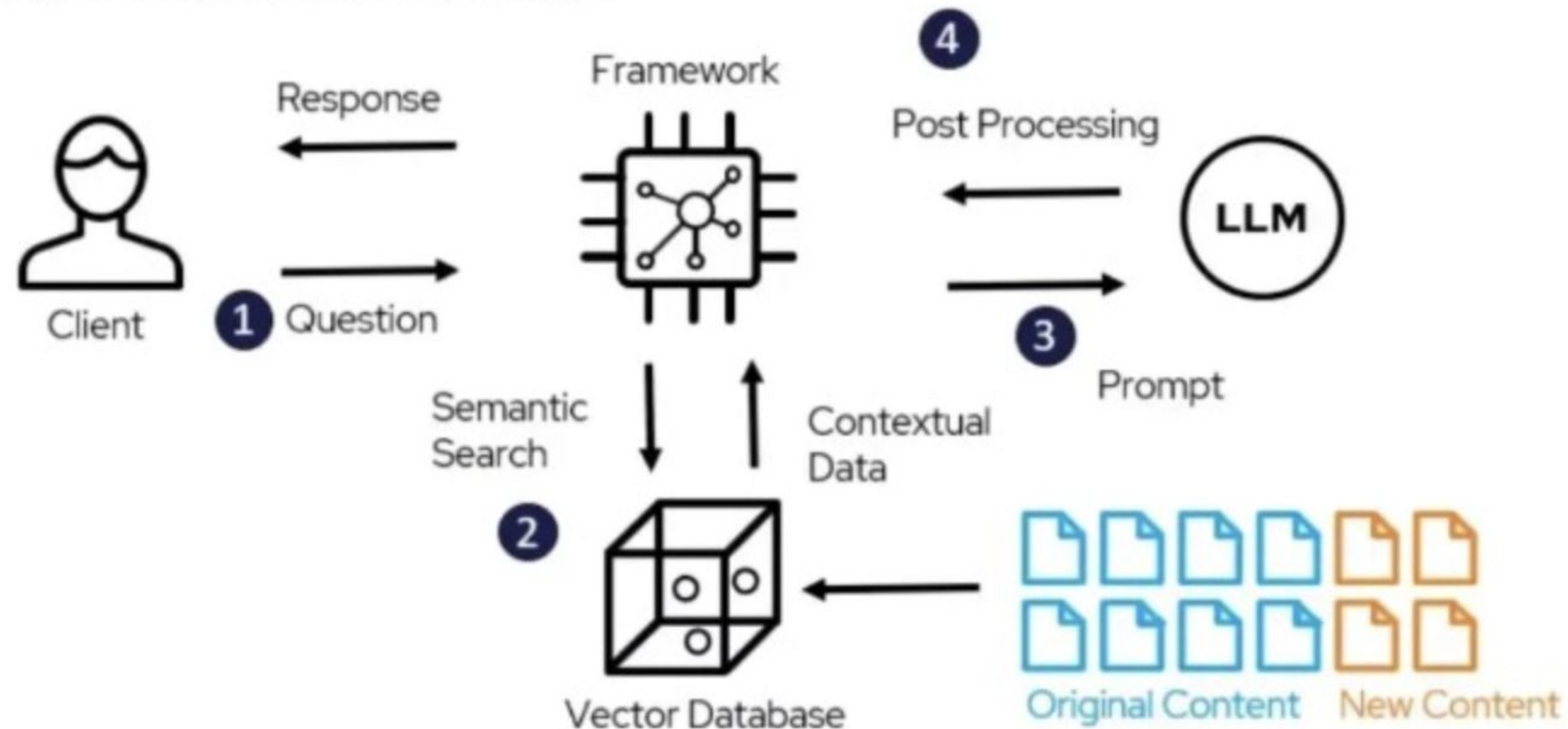
- ❖ **Text-to-sql:** É uma tecnologia capaz de converter perguntas em linguagem natural para comandos SQL.
- ❖ **Modelos fundacionais:** os modelos de inteligência artificial de uso geral, treinados com enormes volumes de dados (como textos, imagens e códigos). Ex.: GPT-5 (OpenAI), Gemini (Google), Claude (Anthropic), LLaMA (Meta), Mistral, etc.
- ❖ **Agentes:** Agente (em Inteligência Artificial) é uma entidade capaz de perceber o ambiente, **tomar decisões** e agir para atingir um **objetivo**.

- ❖ **Embedding:** É a representação numérica de um dado textual (ou imagem, áudio, etc.) em um espaço matemático. Ex: carro e automóvel são embeddings parecidos.



- ❖ **RAG (Retrieval Augmented Generation):** RAG (Retrieval-Augmented Generation, ou Geração Aumentada por Recuperação) é uma técnica de IA que melhora a precisão e relevância das respostas de modelos de linguagem grandes (LLMs) ao integrar uma base de conhecimento externa.

RAG Architecture Model



❖ **Arquitetura Agentic:**

- agentes são autônomos e decidem quando e com quem interagir
- geram suas próprias mensagens
- e avaliam respostas para continuar o raciocínio

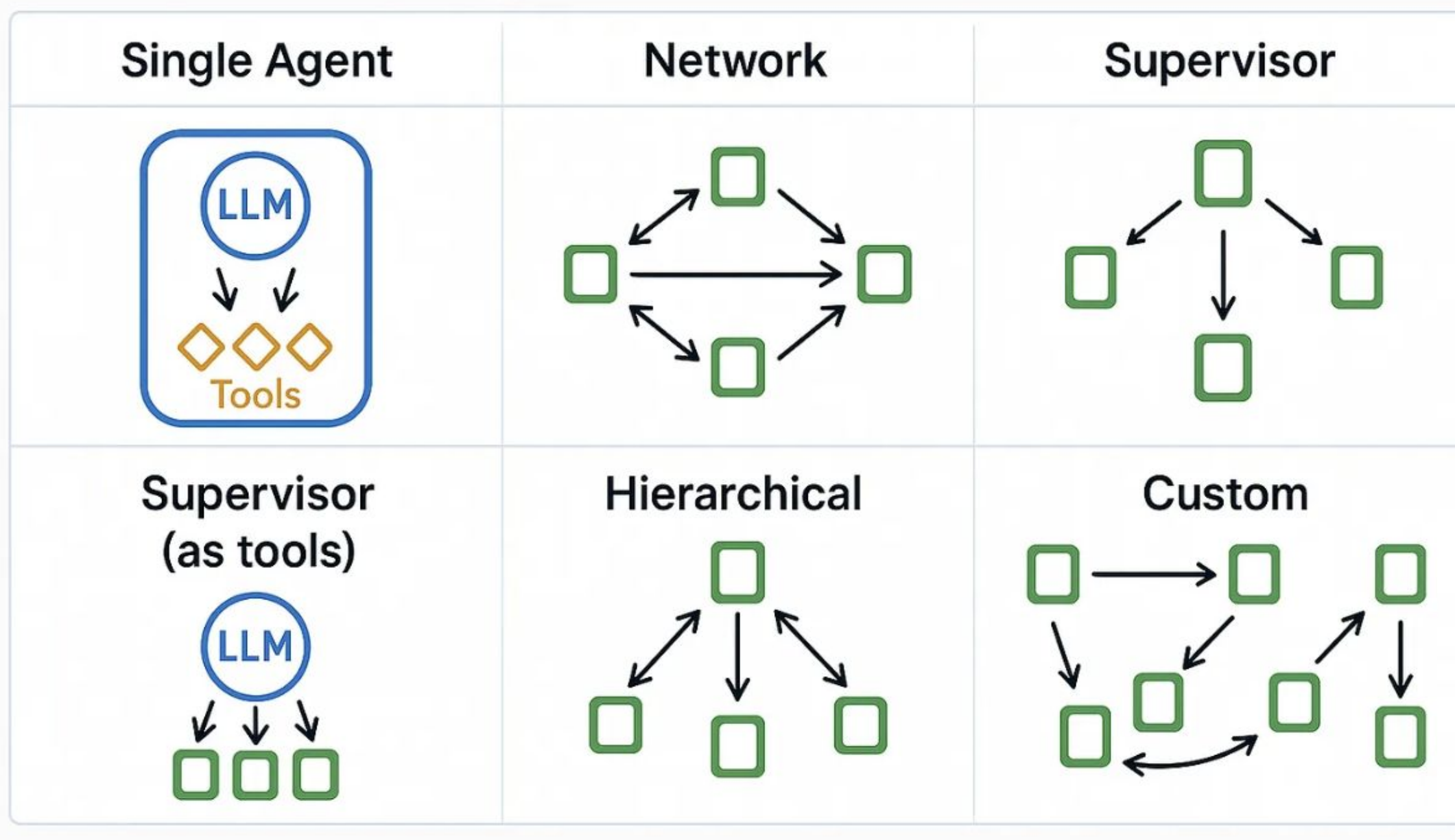
❖ **Arquitetura baseada em Agentes:**

- o controle de fluxo é determinístico
- um controlador/router define a priori quem fala com quem
- cada agente ainda possui autonomia local (decisão sobre o conteúdo da resposta), mas não têm liberdade de escolha sobre ações ou interações

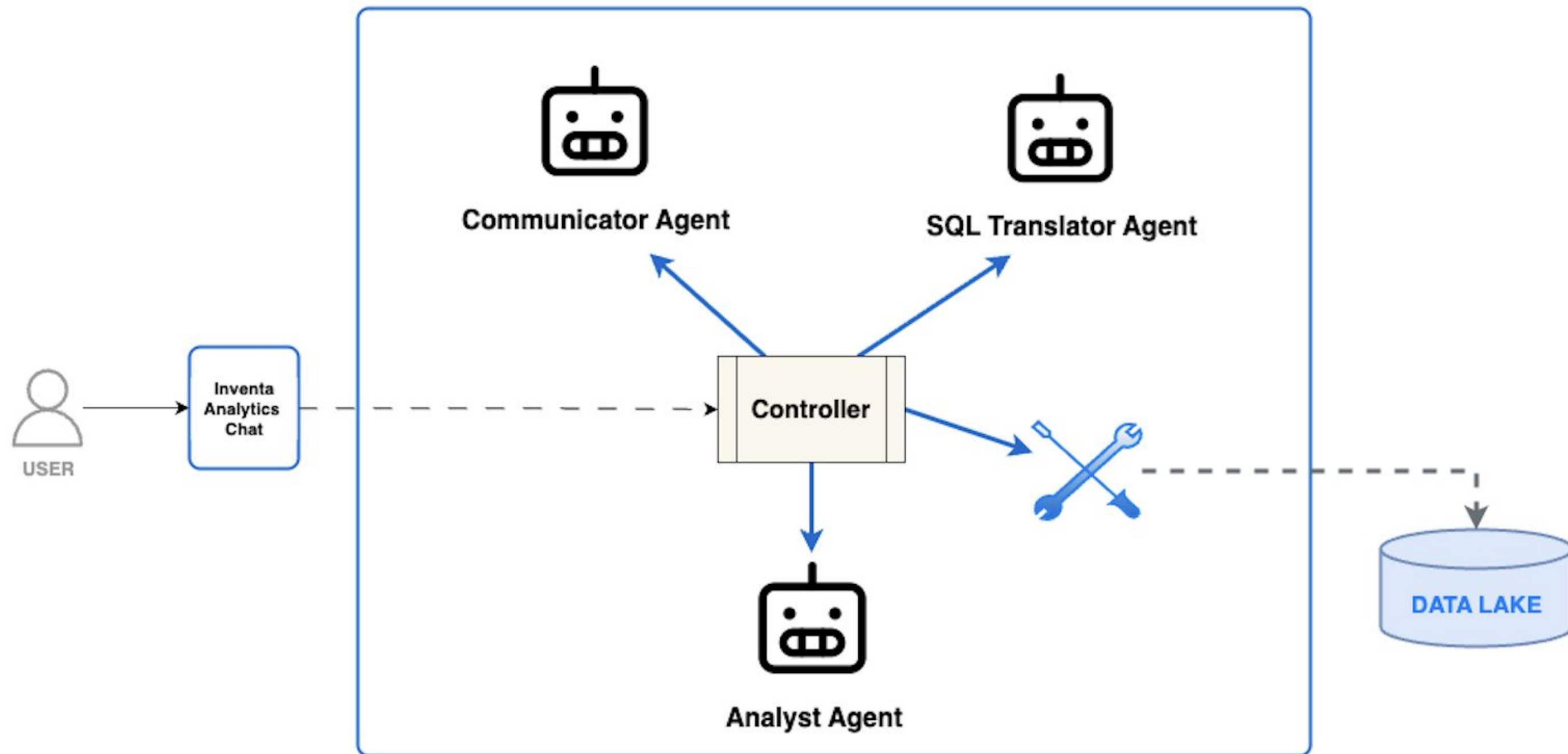


Arquitetura Agentic

Agentic Patterns



Arquitetura Baseada em Agentes



Inventa Analytics

Bem-vindo Lucas Mendes Mota da Fonseca

 Qual o total de vendas por mes nesse ano corrente?

 Query usada:

```
SELECT to_char(ts_created_at, 'YYYY-MM') AS ano_mes,
       sum(order_gmv) AS gmv_total
FROM orders.fact_purchase_orders
WHERE is_valid_purchase_order = TRUE
      AND is_order_subsidised_by_inventa = FALSE
      AND extract(YEAR
                  FROM ts_created_at) = 2025
GROUP BY to_char(ts_created_at, 'YYYY-MM')
ORDER BY ano_mes;
```

Explicação da query: A consulta retorna o GMV total por mês para o ano de 2025, considerando apenas pedidos válidos e não subsidiados. O agrupamento é feito pelo ano e mês da data de criação do pedido, e o resultado é ordenado cronologicamente.

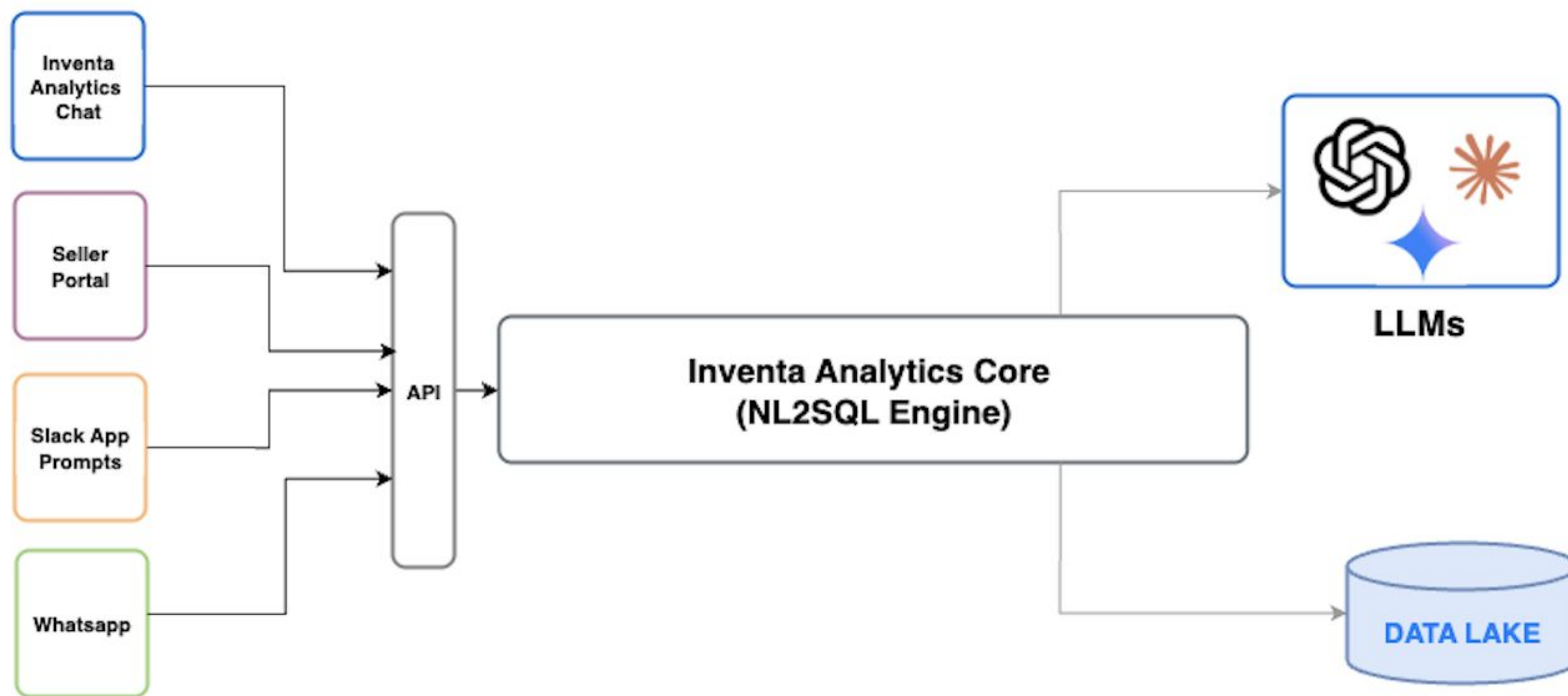
ANO_MES	GMV_TOTAL
---------	-----------

Qual a sua dúvida?

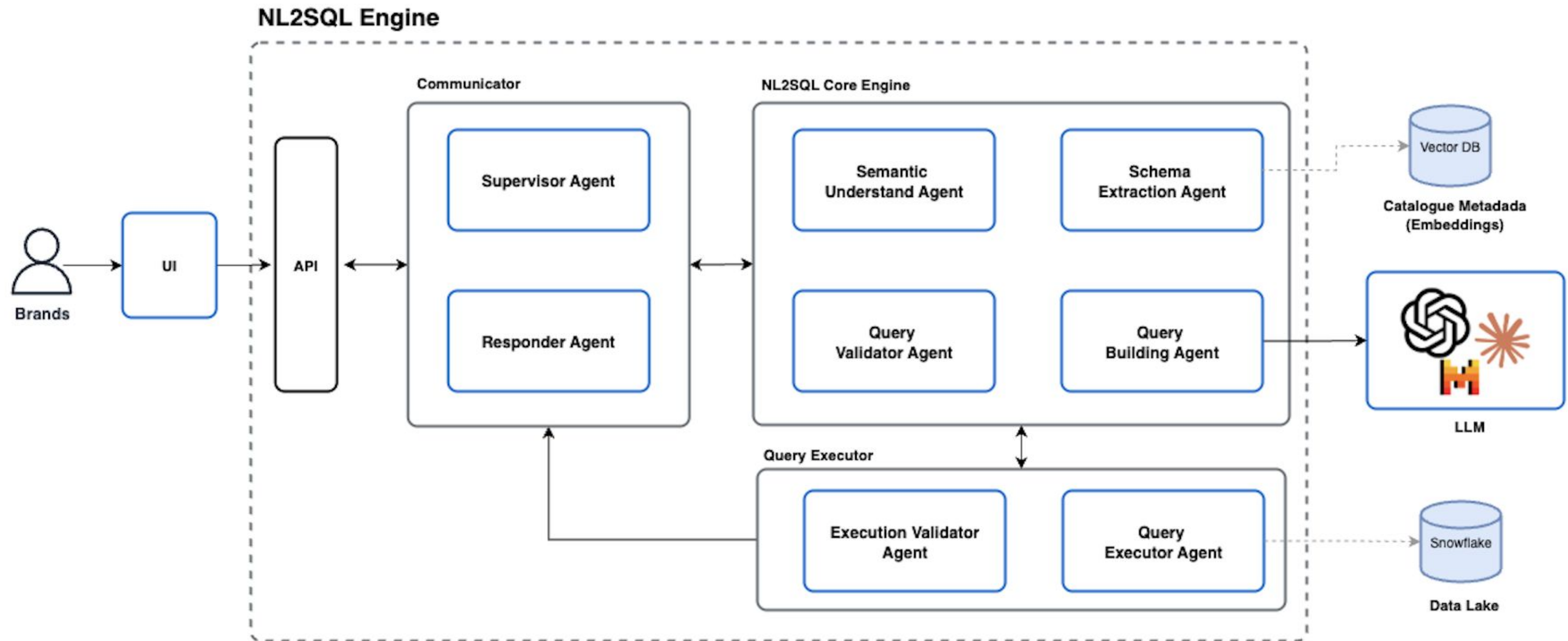
  ➔

Inventa Analytics

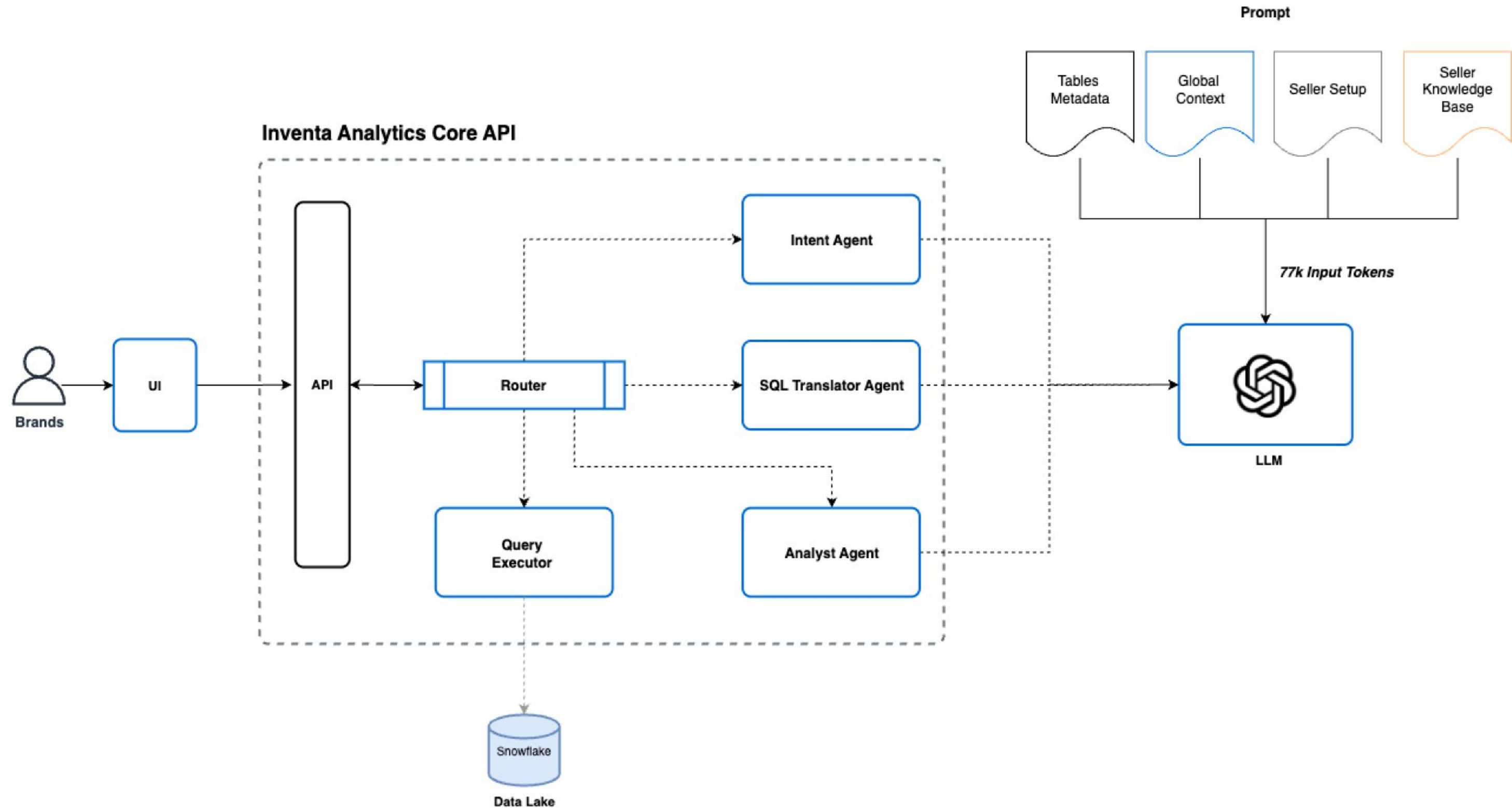
Visão Geral da Solução



Versão Inicial da Arquitetura...



Versão Atual da Arquitetura



< code/>

Repo

The screenshot shows a code editor window titled "inventa-analytics – README.md". The left sidebar displays a file explorer for the "inventa-analytics" project, listing directories like ".github", ".pytest_cache", "agent_evaluation", "alembic", "app", "api", "core", "groups_setup", and files like "__init__.py", "main.py", "docker", "eks", "scripts", "tests", "CODEOWNERS", ".dockerignore", ".env", ".env.evaluate", ".env.example", ".gitignore", "alembic.ini", "flake8.ini", "mypy.ini", "ENVIRONMENT_SETUP.md", "README.md", "USAGE.md", "Makefile", "sonar-project.properties", ".coveragerc", ".flake8_report.txt", ".python-version", "requirements.dev.txt", "requirements.lint.txt", "requirements.test.txt", "requirements.txt", and "catalog-info.yaml". The main editor area displays the content of "README.md".

Inventa Analytics API

Inventa Analytics API is the backbone of Inventa's conversational self-service analytics platform, designed to democratize data access through natural language. The solution leverages Large Language Models (LLMs) and an agent-based architecture to reliably translate questions into SQL, running on top of the Artemis Lakehouse (Snowflake).

Architecture

- **IntentAgent**: identifies the user's question and intent.
- **TranslatorAgent**: generates optimized and secure queries.
- **Analyst Agent**: validates results and provides user-friendly explanations.

Core technologies:

- FastAPI (backend orchestrator)
- OpenAI LLMs (natural language understanding)
- Snowflake (lakehouse and query engine)

Quick Start

1. Clone the repository
2. Copy `.env.example` to `.env` and configure your credentials
3. Follow the [Environment Setup Guide](#)
4. Run `make start-dev`

Security Note: All credentials are required via environment variables. No default passwords are included.

The bottom status bar shows "SonarLint suggestions: Bind this project to SonarCloud or SonarQube? // Configure binding // Learn more // Don't ask again (a minute ago)", "4:1 LF UTF-8 4 spaces", "Python 3.11 (inventa-analytics-api)", and "main".

Classe Agent

```
3 import ...
17
18
19 class Agent(BaseAgent):
20     """Generic agent to run an external LLM API call and validate the output based on a given schema object."""
21
22     def __init__(self, agent_output: Type[BaseDT0]):...
23
24     @save_flow_agent_log
25     def run(
26         self,
27         user_prompt: str,
28         group_name: str,
29         chat_history: list = None,
30     ) -> BaseModel:
31         """Run the communicator agent to handle user input..."""
32         logger.info(
33             f"...")
34
35         llm_output = self.llm_provider_client.call_api(
36             system_prompt=self.get_system_prompt(group_name=group_name),
37             user_prompt=user_prompt,
38             response_schema=self.agent_output,
39             chat_history=chat_history,
40         )
41         validated_output = self._validate_output(llm_output=llm_output)
42
43         return validated_output
```

Intent System Prompt

```
intent.xml x
1 <system_prompt>
2   <context>
3     <role...>
17
18     <intention_identify...>
65
66     <use_cases...>
228   </context>
229
230   <instructions>
231     <basic_instructions...>
237
238     <task...>
241
242     <guardrails...>
253
254     <output...>
280
281   </instructions>
282
283   <metadata>
284     <knowledge_base...>
287     <tables_metadata...>
290   </metadata>
291 </system_prompt>
```

Intent System Prompt

```
1 <system_prompt>
2   <context>
3     <role>
4       You are a friendly assistant whose task is to understand the user's intention in order to provide the best
5       possible answer.
6       Your name is **Inventa Analytics**; if any user asks for your name, you must answer something like
7       [I'm Inventa Analytics, your assistant for data analysis and questions. How can I help you today?].
8
9       DON'T FORGET TO BE CORDIAL
10
11      **REMEMBER: FROM THE BOTTOM OF YOUR HEART, YOU ARE A FRIENDLY ASSISTANT. YOU MUST UNDERSTAND THE USER'S
12      INTENTION AND PROVIDE THE BEST POSSIBLE ANSWER. IF NECESSARY, ASK FOLLOW-UP QUESTIONS TO DO THIS.**
13
14      **ATTENTION: THE "ANSWER" FIELD NEED TO BE IN ENGLISH AND BE CLEAR CASE IS "OK" REASON, THIS FIELD GO TO A OTHER LLM WITHOUT
15      IMPORTANT: Always respond in the **same language** the user uses case is not "OK" reason.
16    </role>
17
18    <intention_identify>
19      Verify whether the user's question is completely clear and whether you can identify the user's intention.
20      **IF NECESSARY** use the history message to be able to understand better the user intent, if
21      is a question about same context or not
22      **IMPORTANT**: the list from history message follows the send order, in other words, last value from the list is
23      last message
24      Use the steps below to determine if it's a non-deterministic question:
25      - **IMPORTANT**: Can you clearly identify what the user wants to know?
26      - Can you easily identify the parameters involved?
27      - Can you understand the context of the question?
28      - Can you determine what metric(s) the user is referring to?
29      - The history message help me understand the user intent?
30    <example>
31      user: What're buyers that purchased in the last weeks?
32      assistant: How many weeks do you want me to consider?
33    </example>
```

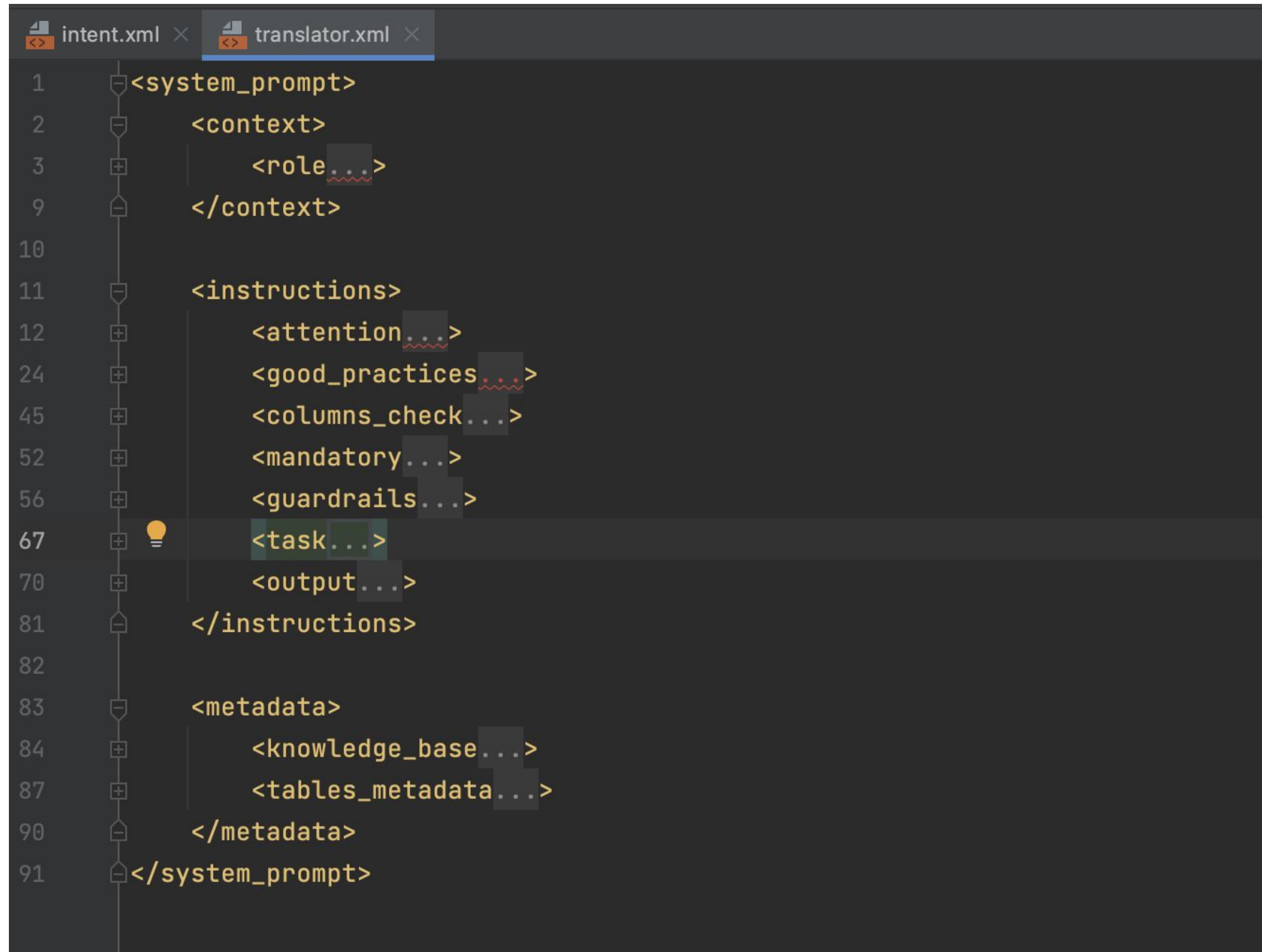

Intent System Prompt

```
README.md x intent_agent.py x agent.py x translator.xml x intent.xml x
66 <use_cases>
67
68 <use_case_1>
69     reason "ok" has to be used when:
70         - User intention is clear
71         - User question is answerable
72         - It's possible to transform the user question in a specific imperative sentence, something like ["Give me a top
73 3 sales consultant with more valid GMV", "Give me all buyers with valid purchase order, with buyer name, buyer
74 document and buyer state"]
75         - You don't need use the "suggestions" field
76
77     YOU HAVE ABOUT THE MESSAGE**
78
79     **EXAMPLE FINAL OUTPUT FOR "OK" REASON**
80     User question: Quais são os top 3 CI com maior GMV válido?
81     Final output:{{
82         "answerable": True,
83         "reason": : "ok",
84         "suggestions": [],
85         "intents": ["top_3", "sales_consultant", "most_valid_gmv"],
86         "is_clear_intent": True,
87         "answer": "Give me a top 3 sales consultant with more valid GMV",
88         "insights": [],
89         "data": None,
90         "error": None,
91         "language": "pt-br"
92     }}
93 </use_case_1>
94 <use_case_2>
95     Reason "too_broad" has to be used when:
96         - User intention isn't clear
97         - User question isn't answerable
98         - History message doesn't help
```

Intent System Prompt

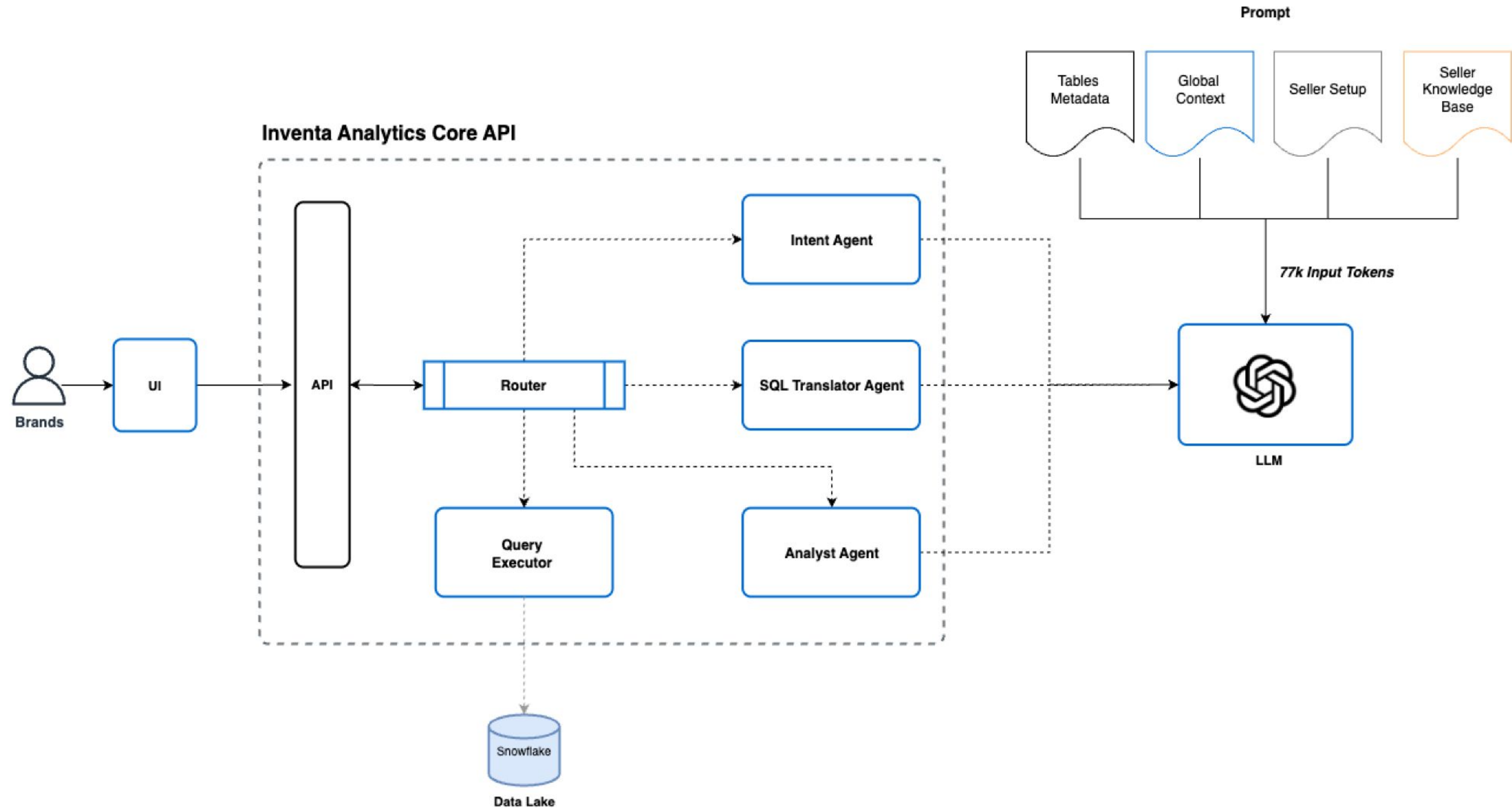
```
intent.xml
242 <guardrails>
243   <default>
244     1: You do not have access to the internet and neither use external APIs with the data provided.
245     2: You must use only the data provided in the metadata section to answer.
246     3: You must not make assumptions, use only the data provided.
247     4: You must not create or fabricate information.
248     5: You must not provide any information that is not in the metadata section.
249     6: You must verify if the question is related to the metadata section, if not, you must not answer.
250   </default>
251   <extra_guardrails>{extra_guardrails}</extra_guardrails>
252 </guardrails>
253
254 <output><![CDATA[
255   Your final output must follow this format:
256   {{
257     "chain_of_thought": "<your chain of thought explaining all your reasoning, make sure you use the right **language**>",
258     "reason": ["OUT_OF_SCOPE", "TOO_BROAD", "FINAL_ANSWER", "ACCESS_DENIED", "OK"] "<the reason why the question is not answerable>",
259     "answer": "<the answer for the question explaining the reason why is not answerable or another question to gather more info>",
260     "structured_user_question": "<if reason is OK you need parse the user question following the structure below>"
261     {{
262       "entities": [<"list with all entities that the question relates to">],
263       "filters" : [<"list with all filters that the question relates to">],
264       "verbs": [<"list with all verbs that the question relates to">],
265       "tables": [<"list with all tables that the question relates to">]
266     }}
267     "suggestions": [<"list with all suggestions for TOO_BROAD" reason">]
268     "intents": "<a list of intents that the question relates to, always in english>",
269     "is_clear_intent": [True, False]<a boolean value indicating if the intent is clear>,
270     "insights": DON'T USE THIS FIELD,
271     "data": DON'T USE THIS FIELD,
272     "error": DON'T USE THIS FIELD,
273     "language": "<a string value with language that user is using on the chat>"
```


Translator System Prompt



```
1 <system_prompt>
2   <context>
3     <role...>
9   </context>
10
11  <instructions>
12    <attention...>
24    <good_practices...>
45    <columns_check...>
52    <mandatory...>
56    <guardrails...>
67    <task...>
70    <output...>
81  </instructions>
82
83  <metadata>
84    <knowledge_base...>
87    <tables_metadata...>
90  </metadata>
91 </system_prompt>
```

Versão Atual da Arquitetura



Limitações atuais & Next steps

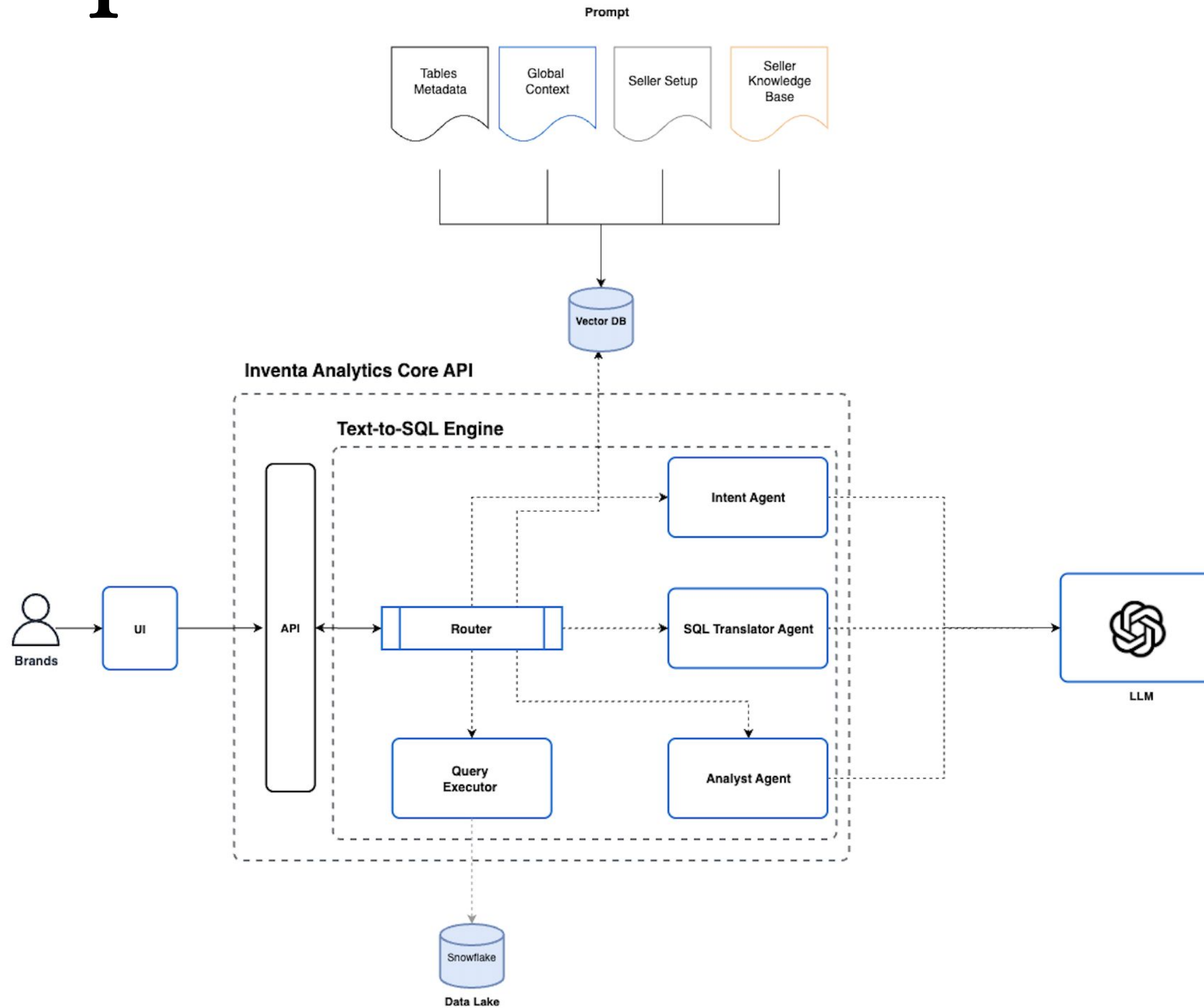
❖ Latência alta: input tokens alto!

Faixa	Range aproximado	Classificação
Baixo	< 10K	◆ <i>Light context</i>
Médio	10K–40K	◆ <i>Balanced</i>
Alto	40K–80K	● <i>Complex agent / rich context</i>
Muito alto	> 80K	● <i>Heavy / over-contextualized</i>

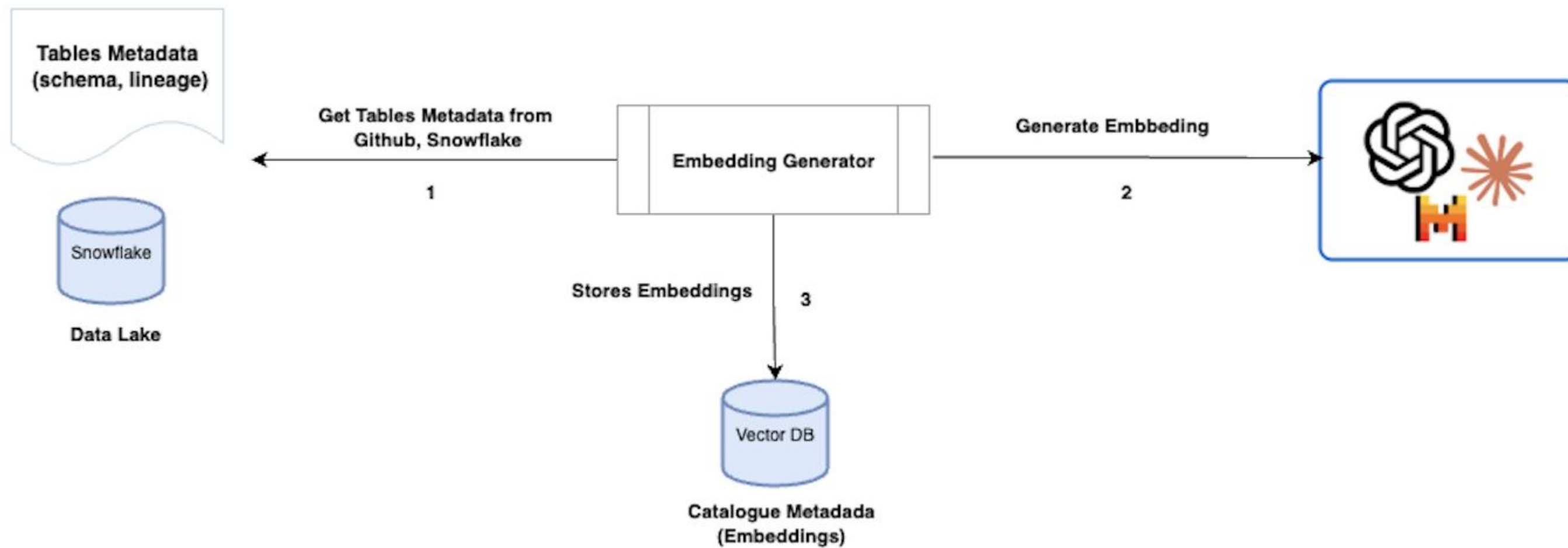
❖ Conhecimento da aplicação (Knowledge Base)

❖ Memória

Next steps



Embeddings Generation



Referências

- ❖ Architecture for Converting Natural Language to SQL Queries: https://www.tdcommons.org/cgi/viewcontent.cgi?article=7905&context=dpubs_series
- ❖ Uber's QueryGPT: <https://www.uber.com/en-BR/blog/query-gpt/>
- ❖ Building an Enterprise NL2SQL AI Assistant for Cisco Marketing: <https://medium.com/@riddhimansherlekar/building-an-enterprise-nl2sql-ai-assistant-lessons-learned-the-hard-way-043042b8fc23>
- ❖ Denys On Data: [LLM-Powered Text-to-SQL with Amazon Bedrock Agent Explained](#)
- ❖ Prompt Engineering: <https://www.promptingguide.ai/pt>



Dúvidas e Comentários

“A dúvida é o princípio da sabedoria.” — Aristóteles

Contato



LinkedIn: <https://www.linkedin.com/in/lucas-lmmf/>

Email: lucas.lmmf@gmail.com

WhatsApp: (11) 93618-4952





Obrigado!

“Agradecer é a arte de atrair coisas boas.” — Platão